Incomplete Angular Time Series Imputation with a Projected Normal Autoregressive Process and Exogenous Predictors

JSM 2025

Benjamin Stockton[1] Ofer Harel[2]

[1]New York University Grossman School of Medicine

[2]University of Connecticut

2025-08-04

Hartford PM2.5 Concentration Data

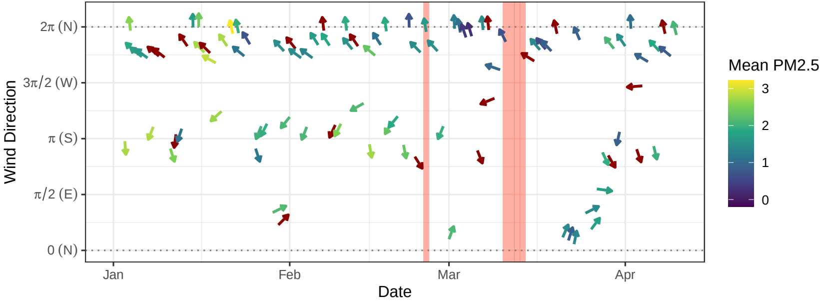


Figure 1: The first 100 days (January 1, 2018 to April 10, 2018) of wind directions. Red bars are where observations are missing.

- Daily data were obtained from the EPA's Air Quality System (AQS) [1] for the period from January 1, 2018 to December 31, 2018.

The Problem

Q: How can we analyze an outcome time series Y with incomplete angular predictors Θ, e.g. air pollution data with meteorological predictors?

Our Proposed Solution

We propose using **multiple imputation** with incomplete angular data imputed by a **projected normal autoregressive process** (PN AR(1)).

Projected Normal Distribution I

Definition

- Let $\mathbf{w}_t \sim N_2(\mu, \Sigma)$ be a latent vector with length $l_t = ||\mathbf{w}_t||$.
- $\mathbf{u}_t = \mathbf{w}_t / l_t = (\cos\theta_t, \sin\theta_t)'$ is a unit vector corresponding to θ_t.
- Then $\theta_t \sim PN(\mu, \Sigma)$.

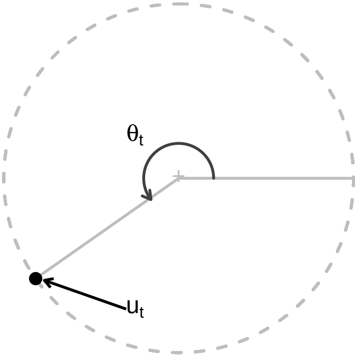


Figure 2: The projected normal angle $\theta_t \sim PN(\mu, \Sigma)$ is equivalent to the unit vector $\mathbf{u}_t$.

Fitting Projected Normal Models

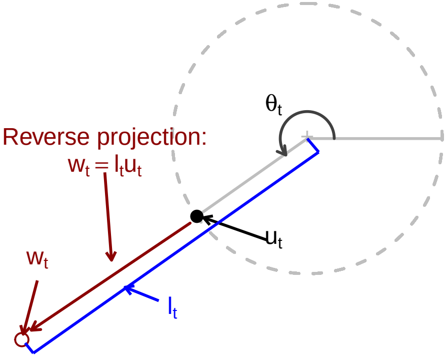


Figure 3: We estimate the latent length l_t to reverse the projection and obtain the latent vector $\mathbf{w}_t = l_t \mathbf{u}_t \sim N_2(\mu, \Sigma)$. Repeat with the whole data set to estimate μ and Σ using $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)'$ where $\mathbf{w}_t \overset{iid}{\sim} N_2(\mu, \Sigma)$.

Multiple Imputation Procedure [2, 3]

1. Multiply Impute M times with $\dot{\theta}_{imp}^{(m)} = g(\mathbf{X}, \Theta_{obs})$.

Incomplete Data: $(\Theta_{obs}, \Theta_{mis}, \mathbf{Z})$

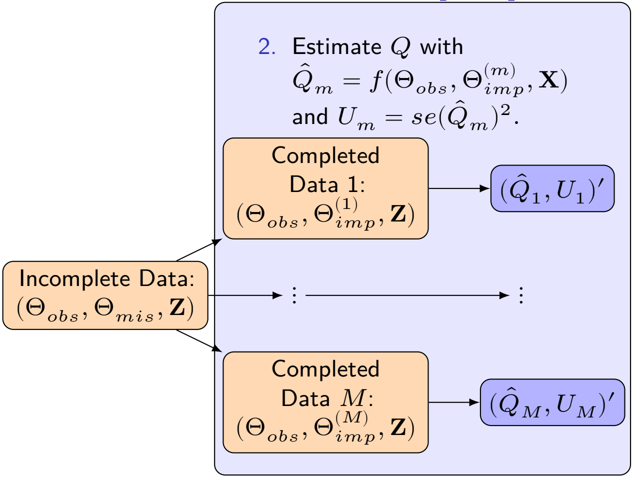
Completed Data 1: $(\Theta_{obs}, \Theta_{imp}^{(1)}, \mathbf{Z})$

$\vdots$

Completed Data M: $(\Theta_{obs}, \Theta_{imp}^{(M)}, \mathbf{Z})$

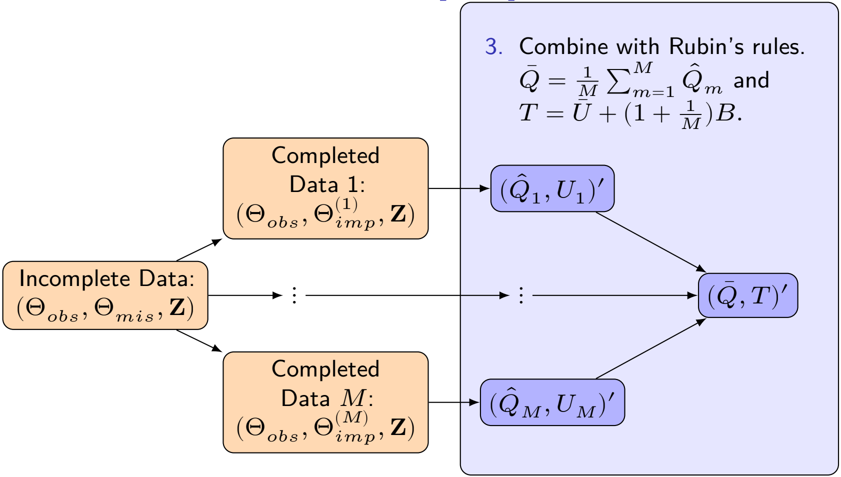
$\Theta = (\theta_0, \dots, \theta_T)'$ and $\mathbf{Z} = (Y, \mathbf{X})'$

Multiple Imputation Procedure [2, 3]



Multiple Imputation Procedure [2, 3]

3. Combine with Rubin's rules.
$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M} \hat{Q}_m$ and
$T = \bar{U} + (1 + \frac{1}{M})B$.



$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m - \bar{Q})^2 \text{ and } \bar{U} = \frac{1}{M}\sum_{m=1}^{M} U_m.$$

Projected Normal Autoregressive Process

PN AR(1) Process

- Let $\theta_t| \ . \ \sim PN(\mu_t, \Sigma)$ where

$$\mu_t = \mu_0 + \Phi l_t \mathbf{u}_{t-1} + \mathbf{B}\mathbf{x}_t$$

- l_t are latent lengths such that $W_t = l_t(\cos\theta_t, \sin\theta_t)' \sim N_2(\mu_t, I_2)$.
- Weakly informative normal priors for $l_t \sim N_+(0, 10^2)$, $\beta_{ik} \sim N(0, 100^2)$ and ϕ_{ij} $N_{[-1,1]}(0, 100^2)$.

- The PN AR(1) process [4] is implemented in the Bayesian setting to draw imputations from the posterior predictive distribution.

Projected Normal-based Imputation

Imputation Procedure

1. **Fit** the PN AR(1) model using the observed data $(\Theta_{obs}, \mathbf{X}, Y)$ and obtain posterior draws for the model parameters $(\mathbf{B}^{(q)}, \Phi^{(q)})'$ for the last MCMC iteration q.

Projected Normal-based Imputation

Imputation Procedure

1. **Fit** the PN AR(1) model using the observed data $(\Theta_{obs}, \mathbf{X}, Y)$ and obtain posterior draws for the model parameters $(\mathbf{B}^{(q)}, \Phi^{(q)})'$ for the last MCMC iteration q.
2. For each missing observation, **sample** $\dot{\mathbf{w}}_t \sim N_2(\dot{\mu}_t^{(q)}, I_2)$ and **project** onto the unit circle to get $\dot{\mathbf{u}}_t = \dot{\mathbf{w}}_t / ||\dot{\mathbf{w}}_t||$.

Projected Normal-based Imputation

Imputation Procedure

1. **Fit** the PN AR(1) model using the observed data $(\Theta_{obs}, \mathbf{X}, Y)$ and obtain posterior draws for the model parameters $(\mathbf{B}^{(q)}, \Phi^{(q)})'$ for the last MCMC iteration q.
2. For each missing observation, **sample** $\dot{\mathbf{w}}_t \sim N_2(\dot{\mu}_t^{(q)}, I_2)$ and **project** onto the unit circle to get $\dot{\mathbf{u}}_t = \dot{\mathbf{w}}_t / ||\dot{\mathbf{w}}_t||$.
3. **Convert** $\dot{\mathbf{u}}_t$ to $\dot{\theta}_t$ and use these posterior predictive draws to **impute** the missing θ_t.

Projected Normal-based Imputation

Imputation Procedure

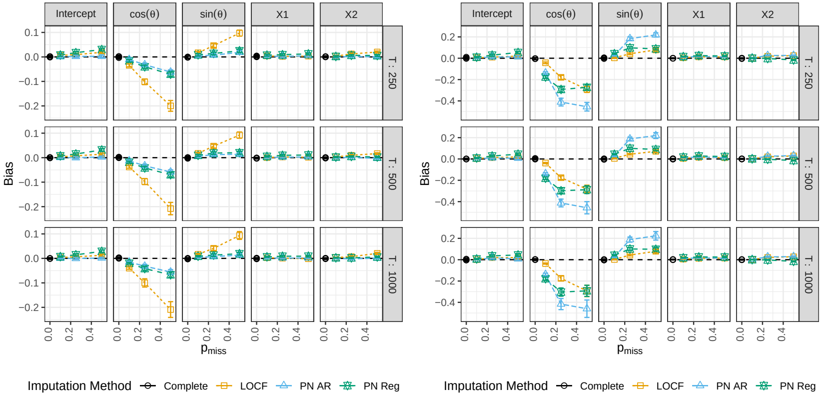
1. **Fit** the PN AR(1) model using the observed data $(\Theta_{obs}, \mathbf{X}, Y)$ and obtain posterior draws for the model parameters $(\mathbf{B}^{(q)}, \Phi^{(q)})'$ for the last MCMC iteration q.
2. For each missing observation, **sample** $\dot{\mathbf{w}}_t \sim N_2(\dot{\mu}_t^{(q)}, I_2)$ and **project** onto the unit circle to get $\dot{\mathbf{u}}_t = \dot{\mathbf{w}}_t / ||\dot{\mathbf{w}}_t||$.
3. **Convert** $\dot{\mathbf{u}}_t$ to $\dot{\theta}_t$ and use these posterior predictive draws to **impute** the missing θ_t.
4. **Repeat** M times, refitting the model each time.

Set-Up

- **Aims** - Evaluate the LOCF [5], PN Regression [6, 7], and PN AR(1) imputation models with an ARX analysis model.
- **Data Generating Mechanism** - Simulate response variable from a ARX model and angular data from PN AR(1) model.
  - Simulate with *high* or *low* autocorrelation, varied sample sizes, and varied proportions of missingness.
- **Estimand** - The regression coefficient vector β.
  - $Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 \cos\theta_t + \beta_4 \sin\theta_t + \phi_y Y_{t-1} + \epsilon_t$
  - $\epsilon_t \overset{iid}{\sim} N(0, \sigma_y^2)$
- **Performance Measure** - Bias and 95% CI Coverage[1]

[1] The code for the simulations is available at Github: https://github.com/benjamin-stockton/ch2-ts-mi-sim.

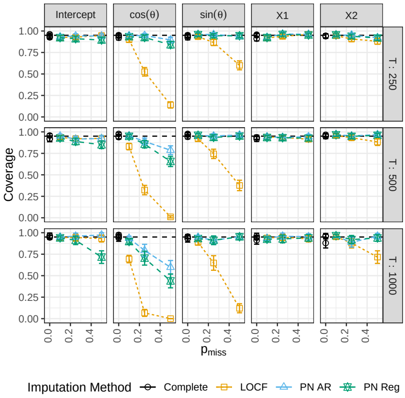
Results



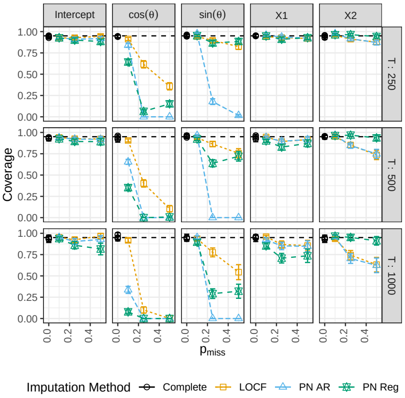
(a) Low Autocorrelation

(b) High Autocorrelation

Figure 4: Regression coefficient bias.



(a) Low Autocorrelation



(b) High Autocorrelation

Figure 5: 95% confidence interval coverage. The horizontal dashed line is at 95% coverage.

Hartford PM2.5 Concentration Data

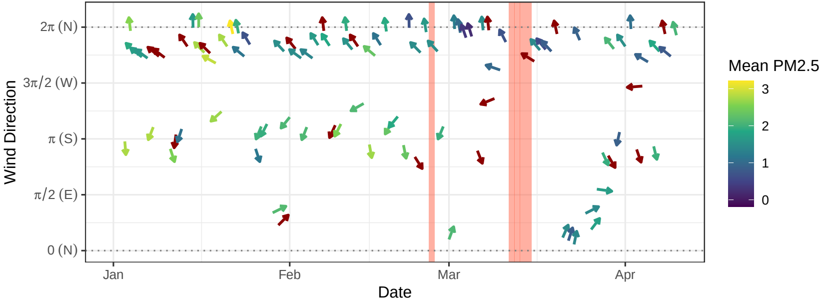


Figure 6: The first 100 days (January 1, 2018 to April 10, 2018) of wind directions. Red bars are where observations are missing.

Imputed Wind Directions

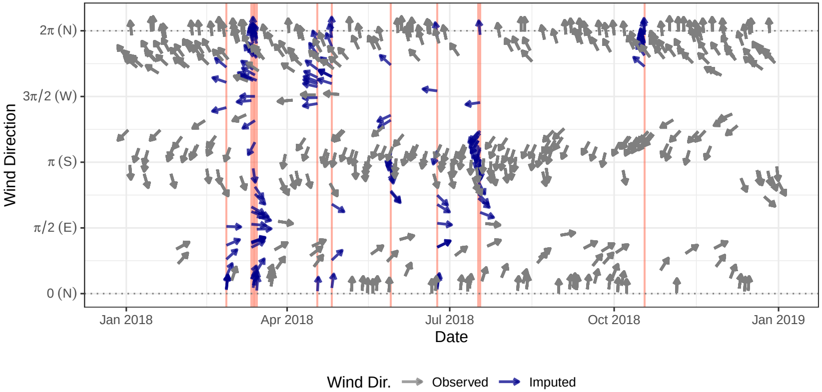


Figure 7: Ten (out of 25 total) completed data sets imputed with PN AR(1) and predictive mean matching using MICE [8]. 3% of wind observations and 13% of PM2.5 observations are missing.

Results

Table 1: Parameter estimates from the multiply imputed PM2.5 and wind data.

Variable	Est.	95% LB	95% UB
Intercept	6.55	5.82	7.29
Max WindSpeed	-1.05	-1.62	-0.49
$\cos\theta$	0.14	-0.41	0.68
$\sin\theta$	-0.73	-1.73	0.29
Max WindSpeed $\times \cos\theta$	0.47	-0.15	1.10
Max WindSpeed $\times \sin\theta$	-0.46	-1.69	0.77
$\cos\theta \times \sin\theta$	-0.22	-1.64	1.19
Max WindSpeed $\times \cos\theta \times \sin\theta$	1.46	-0.32	3.27
ϕ_1	0.49	0.37	0.59
σ_Y	3.33	3.07	3.63

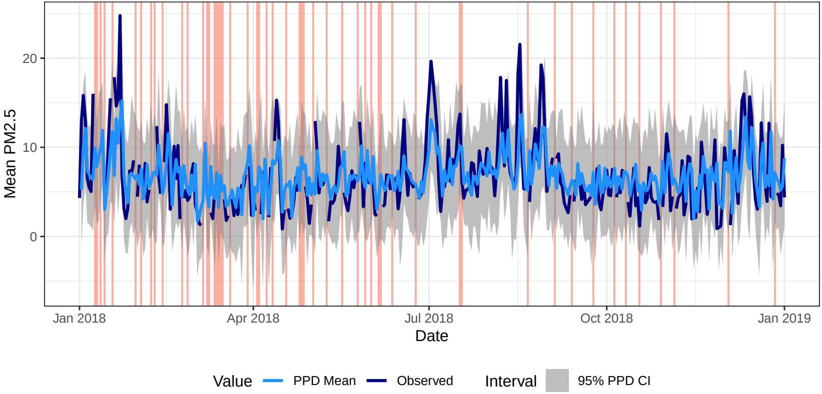


Figure 8: Posterior predictive mean PM2.5 concentration with 95% CI and observed data.

Conclusion

- We have developed a projected normal autoregressive model for imputing angular time series with MICE [8].[2]
- The proposed method works best under low-to-moderate autocorrelation settings. LOCF may be a viable alternative with high autocorrelation.
- Multiply imputed regression analysis of daily PM2.5 concentrations in Hartford, CT showed no to weak associations between lagged maximum daily wind speed and direction and the PM2.5 concentration.

[2]Available at: https://github.com/benjamin-stockton/pnregstan and https://github.com/benjamin-stockton/imputeangles

Acknowledgements

- Research was supported by NSF AGEP-GRS supplement for Award #2015320.
- The computational work performed on this project was done in part on the Storrs High-Performance Computing cluster. We would like to thank the UConn Storrs HPC, NYU Big Purple, and both HPC technical support teams for providing the resources and support that contributed to these results.

References I

1. US Environmental Protection Agency, O. (2021, August). Air Quality System Data Mart [internet database]. Data and Tools, https://www.epa.gov/outdoor-air-quality-data.

2. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys | Wiley Series in Probability and Statistics*. New York: Wiley.

3. Hopke, P. K., Liu, C., & Rubin, D. B. (2001). Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic. *Biometrics*, *57*(1), 22–33. Retrieved from https://www.jstor.org/stable/2676838

4. Fisher, N. I., & Lee, A. J. (1994). Time Series Analysis of Circular Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(2), 327–339. Retrieved from https://www.jstor.org/stable/2345903

5. Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207. https://doi.org/10.32614/RJ-2017-009

6. Nuñez-Antonio, G., Gutiérrez-Peña, E., & Escarela, G. (2011). A Bayesian regression model for circular data based on the projected normal distribution. *Statistical Modelling*, *11*(3), 185–201. https://doi.org/10.1177/1471082X1001100301

7. Hernandez-Stumpfhauser, D., Breidt, F. J., & Woerd, M. J. van der. (2017). The General Projected Normal Distribution of Arbitrary Dimension: Modeling and Bayesian Inference. *Bayesian Analysis*, *12*(1), 113–133. https://doi.org/10.1214/15-BA989

8. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Appendix

Table 2: Summary of missing data proportions for each variable.

Variable	n_{miss}	p_{miss}
PM2.5	47	0.128
Pressure	8	0.022
Relative Humidity	58	0.158
Temperature	10	0.027
Wind Direction	12	0.033
Wind Speed	12	0.033